

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(6): 23-29
© 2024 Stats & Maths
<https://www.mathsjournal.com>
Received: 03-09-2024
Accepted: 06-10-2024

Rohit Kumar Verma
Associate Professor, Department
of Mathematics, Bharti
Vishwavidyalaya, Durg,
Chhattisgarh, India

Jharana Chandrakar
Ph.D., Research Scholar,
Department of Mathematics,
Bharti Vishwavidyalaya, Durg,
Chhattisgarh, India

Corresponding Author:
Rohit Kumar Verma
Associate Professor, Department
of Mathematics, Bharti
Vishwavidyalaya, Durg,
Chhattisgarh, India

Uncertainty estimation of mutual information with joint distribution of marginal random variables

Rohit Kumar Verma and Jharana Chandrakar

DOI: <https://dx.doi.org/10.22271/math.2024.v9.i6a.1881>

Abstract

In this proposal, we suggest using marginal probabilities to determine the joint distribution of two random variables. Additionally, we provide a method for calculating the uncertainty of mutual information in order to determine mutual information in machine learning. Lastly, we apply our newly discovered mutual knowledge to credit scoring.

Keywords: Mutual information, probability theory, joint entropy, entropy, conditional entropy, machine learning etc.

Introduction

In recent times, mutual information has become a crucial indicator of statistical reliance. In engineering, particularly in machine learning, it has been applied as a criterion for feature selection (see ^[1-3] and the references therein). Shannon originally presented its antecedent, known as the rate of transmission, in a classic work ^[4] for the communication system in 1948. Entropy was initially established by Shannon for a single discrete chance variable. Using the joint distribution, he subsequently established the joint entropy and conditional entropy for two discrete chance variables. Lastly, he stated that the difference between the entropy and conditional entropy represents the rate of transmission. Different definitions of mutual information have been proposed since Shannon's groundbreaking work ^[4]. These definitions can be broadly classified into two classes: (1) definitions involving random variables and (2) definitions involving ensembles, or in mathematical literature, probability spaces.

The joint distribution of two random variables is necessary for the definition of mutual information. More specifically, Kullback (1959) ^[5] defined entropy, conditional entropy, and joint entropy using elementary mathematical techniques. Using probability theory, Pinsker (1960, and 1964) ^[6] tackled Shannon's basic ^[15, 16] ideas in a more sophisticated way. His definition of mutual information was more expansive, implying that the two random variables had different probability domains.

Using Shannon's method, Ash (1965) ^[7] defined mutual information, specifically presuming that the two random variables shared the same probability space ^[17].

The joint probability measure of the joint sample space of two ensembles determines the definitions. Three authors of such definitions—Fano (1961) ^[9], Abramson (1963) ^[10], and Gallager (1968) ^[11]—evolved their concept in a similar way. They defined condition entropy, ensemble entropy, and joint entropy of two ensembles first.

They then clarified what a joint event's mutual information was. They determined the mean value of the mutual information of a joint event, noting that it is a random variable. They dubbed the outcome the mean information of two ensembles. The definition of y's conditional entropy $H_x(y)$ is

$$H_x(y) = -\sum_{i,j} p(i,j) \log p(j/i) = -\sum_{i,j} p(i,j) \log \frac{p(i,j)}{\sum_j p(i,j)} \quad (1.1)$$

Then the following relation holds

$$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x)$$

$$H(x) - H_y(x) = H(y) - H_x(y) = H(x) + H(y) - H(x, y) \quad (1.2)$$

The difference between $H(x)$ and $H_y(x)$ is the definition of the information transmission rate, or R . As a result, there are two methods to express R :

$$R = H(x) - H_y(x) = H(y) - H_x(y) = H(x) + H(y) - H(x, y) \quad (1.3)$$

Remark 11. Shannon did not discover the precise equation for R :

$$R = \sum_{i,j} p(i, j) \log \frac{p(i, j)}{p(i)p(j)} \quad (1.4)$$

Give the following definitions for the joint, entropy, and conditional entropies:

$$H(x, y) = - \sum_i \sum_j p_{ij} \log p_{ij}$$

$$H(x) = - \sum_i p_i \log p_i$$

$$H(y) = - \sum_j p_j \log p_j$$

$$H(y/x) = \sum_i p_i H(y/x_i) = - \sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i} \quad (1.5)$$

$$H(y/x) = \sum_i p_i H(y/x_i) = - \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i}$$

$$\text{Then } H(x, y) = H(x) + H(y/x) \leq H(x) + H(y) \text{ and } H(y) \geq H(y/x)$$

Ash^[7] started with two arbitrary variables. X and Y made the assumption that their probability spaces were identical. By using Shannon's approach, he methodically described the entropy, conditional entropy, and joint entropy in^[4]. Finally, he indicated $H(X) - H(X/Y)$ by $I(X/Y)$ and referred to it as the data that Y had sent regarding X . Pinsker^[6] used probability theory to address Shannon's basic ideas in a more sophisticated way. Assume that there are two random variables δ and μ : one defined on a probability space $(\Omega, S_\omega, P_\delta)$, and the other one defined on a probability space (Ψ, S_ϕ, P_μ) , and both of them are taking values in a measurable space (X, S_x) and (Y, S_y) . Then, the pair δ, μ of random variables may be regarded as a single random variable (δ, μ) with values in the product space $X \times Y$ of all pairs (x, y) with $x \in X, y \in Y$.

The distribution of $(\cdot)$ is referred to as the joint distribution of random variables δ and μ . The product of the distributions $P_\delta(\cdot)$ and $P_\mu(\cdot)$ represents the distribution $P_{\delta\mu}(\cdot)$ defined on $S_x \times S_y$.

$$P_{\delta\mu}(EXF) = P_\delta(E)P_\mu(F) \quad (1.6)$$

For $E \in S_x$ and $F \in S_y$

If X and Y are discrete random variables with countably many points each, for example, $\{x_1, x_2, \dots\}$ and $\{y_1, y_2, \dots\}$ then

$$I(\delta, \mu) = \sum_{i,j} P_{\delta\mu}(x_i, y_j) \log \frac{P_{\delta\mu}(x_i, y_j)}{P_\delta(x_i)P_\mu(y_j)} \quad (1.7)$$

$I(\delta, \mu)$ is called the information of δ and μ with respect to the other.

Assume X is a discrete variable with the probability mass function $P(x) = \Pr\{X=x\}$, $x \in X$ and the alphabet X . The probability mass function $p(y) = \Pr\{Y=y\}$, $y \in Y$ and the discrete variable Y with the alphabet Y are both assumed.

Suppose that the joint mass function (joint distribution) of X and Y is $p(x, y)$. The mutual information $I(X, Y)$ can then be defined as

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

Gallager^[11] later introduced the idea of mutual information into communication theory using a more complete and in-depth method. In fact, he merged and combined the findings from Abramson^[10] and Fano^[9].

Let's assume that each discrete ensemble, X and Y , has a sample space $\{a_1, a_2, \dots, a_k\}$ and $\{b_1, b_2, \dots, b_j\}$

Consider the joint sample space $\{(a_k, b_j): 1 \leq k \leq K, 1 \leq j \leq J\}$. A probability measure on the joint sample space is given by the joint probability $P_{XY}(a_k, b_j)$, defined for $1 \leq k \leq K, 1 \leq j \leq J$.

A joint sample space multiplied by a joint probability measure for outcomes x and y yields a joint XY ensemble.

It is thus possible to determine the marginal probability as

$$P_X(a_k) = \sum_{j=1}^J P_{XY}(a_k, b_j), k = 1, 2, \dots, K \quad (1.8)$$

$$P_Y(b_j) = \sum_{k=1}^K P_{XY}(a_k, b_j), j = 1, 2, \dots, J \quad (1.9)$$

This is expressed in shorter notation as

$$P(y) = \sum_x P(x, y) \quad (1.10)$$

Given that outcome a_k of x is, the conditional probability that outcome y is b_j , if $P_X(a_k) > 0$, is defined as

$$P_{(Y/X)}(b_j/a_k) = \frac{P_{XY}(a_k, b_j)}{P_X(a_k)} \quad (1.11)$$

The definition of the mutual information between occurrences $x = a_k$ and $y = b_j$ is

$$\begin{aligned} I_{X:Y}(a_k, b_j) &= \log \frac{P_{(Y/X)}(b_j/a_k)}{P_X(a_k)} = \log \frac{P_{XY}(a_k, b_j)}{P_X(a_k)P_Y(b_j)} \\ &= \log \frac{P_{XY}(a_k, b_j)}{P_X(a_k)} = I_{Y:X}(a_k, b_j) \end{aligned} \quad (1.12)$$

$$I(X, Y) = \sum_{k=1}^K \sum_{j=1}^J P_{XY}(a_k, b_j) \log \frac{P_{XY}(a_k, b_j)}{P_X(a_k)P_Y(b_j)} \quad (1.13)$$

By means of an information channel consisting of a transmitter of alphabet A with elements a_i and total elements t and a receiver of alphabet B with elements b_j and total elements r , Abramson^[10] denoted $H(A) - H(A/B) = \sum_{A,B} P(a, b) \log(P(a, b)/P(a)P(b))$ by $I(A, B)$ and called it mutual information of A and B . The rate of transmission in^[1] or the mutual information $I(X, Y)$ between two continuous random variables X and Y ^[8] is defined as

$$I(X, Y) = \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy, \quad (1.14)$$

However, differential mutual information is much less prevalent than discrete mutual information. On the one hand, the joint density function in question needs to be calculated because it is typically unknown^[13, 14]. However, as most data utilised in machine learning and engineering are finite, mutual information between discrete random variables is used.

Our Claims

In information theory Kullback's formulation was more concise and mathematical. While Ash's^[12] definition adhered to Shannon's approach, it was more methodical. Because Pinsker's formulation made use of probability theory, it was the most mathematical. In terms of communication theory, Gallager's definition was more thorough and comprehensive. Because of its brevity, Cover and Thomas's concept has become accepted as standard practice in information theory.

Some mathematical errors exist in these different definitions of mutual information, though. Marginal probabilities are redefined from the joint probabilities. In actuality, the marginal probabilities need not be redefined from the joint probabilities since they are provided by the ensembles. With the exception of Pinsker's definition, either assume that the two random variables have the same probability space or ignore the probability spaces. The above definitions make the assumption that there is a joint probability measure or distribution. However, they all fail to acknowledge a crucial fact: the joint probability measure or joint distribution is not unique.

Now, if X be a finite discrete random variable on discrete probability space (Ω_1, F_1, P_1) with $\Omega_1 = \{\omega_1, \omega_2, \dots\}$. and range $\{x_1, x_2, \dots\}$ with $K \leq n$. Let Y be a discrete random variable on probability space (Ω_2, F_2, P_2) with $\Omega_2 = \{\omega_1, \omega_2, \dots\}$. and range $\{y_1, y_2, \dots, y_L\}$. with $L \leq m$.

If X and Y have the same probability space (Ω, F, P) , then the joint distribution is simply

$$P_{XY}(X = x, Y = y) = P(\omega \in \Omega: X(\omega) = x, Y(\omega) = y) \quad (2.1)$$

The joint distribution becomes more complex, though, when X and Y have separate probability spaces and, consequently, different probability measures.

A joint probability measure of P_{XY} of P_1 and P_2 is

$$P_1(A) = P_{XY}(A/\Omega_2) = \sum_{j=1}^m P_{XY}(A/\{p_j\}) \quad (2.2)$$

$$P_2(B) = P_{XY}(\Omega_1/B) = \sum_{i=1}^m P_{XY}(\{\omega_i\}/B)$$

Claim 2.1. The product probability measure, also known as the product distribution, is the joint probability measure that exists for any two discrete random variables.

Proof: As previously stated, let X and Y be random variables. Define a function from $(\Omega_1 \times \Omega_2)$ to $[0, 1]$ as follows:

$$P_{XY}(\omega_i, p_j) = P_1(\omega_i)/P_2(p_j) \quad (2.3)$$

$$\text{Then } \sum_{i=1}^n \sum_{j=1}^m P_{XY}(\omega_i, p_j) = \sum_{i=1}^n \sum_{j=1}^m P_1(\omega_i)/P_2(p_j)$$

$$= \sum_{i=1}^n P_1(\omega_i) / \sum_{j=1}^m P_2(p_j) = 1 \quad (2.4)$$

Hence, P_{XY} can serve as a probability measure on $F_1 \times F_2$ by Definition 2. The probability of any event $A \times B \subseteq \Omega_1 \times \Omega_2$ is computed simply by adding the probabilities of the individual points of $(\omega, p) \in A \times B$. Moreover, for any $A = \{\omega_{i_1}, \omega_{i_2}, \dots, \omega_{i_s}\} \in \Omega_1$ of s elements,

$$P_{XY}(A/\Omega_2) = \sum_{j=1}^m P_{XY}(A/\{p_j\})$$

$$= \sum_{j=1}^m \sum_{u=1}^s P_{XY}(\{\omega_{i_u}\}/\{p_j\})$$

$$= \sum_{j=1}^m P_2(p_j) / \sum_{u=1}^s P_1(\omega_{i_u})$$

$$= \sum_{u=1}^s P_1(\omega_{i_u}) = P_1(A) \quad (2.5)$$

Similarly, $P_{XY}(\Omega_1/B) = P_2(B)$ for any $B \in \Omega_2$. Hence, $P_{XY}(\{X = x_i\}/\{Y = y_j\}) = P_1(\{X = x_i\}) \text{ or } P_2(\{Y = y_j\})$ is a combined X and Y probability measure (see) ^[8]. We study the discrete probability distributions P and Q.

Then $D(P // Q) \geq 0$ with equality if and only if $p_i = q_i$ for all i. Since the Kullback-Leibler distance is not symmetric and does not meet the triangle inequality, it cannot be considered a real distance between distributions. However, it is frequently helpful to consider relative entropy to be a "distance" between distributions. The subsequent characteristic demonstrates that the Kullback-Leibler distance between the joint distribution P_{XY} and the product distribution $P_X P_Y$ represents mutual information under a joint probability measure.

Claim 2.2: The Kullback-Leibler distance between the joint distribution and the product distribution is the mutual information of random variables X and Y.

Proof. Through the use of a mapping from two-dimensional to one-dimensional indices, $(i, j) \rightarrow (i - 1) * L + j \triangleq n$

$$\text{for } i = 1, \dots, K, j = 1, 2, \dots, L, \quad (2.6)$$

and making use of an additional mapping that converts one-dimensional indices back to two-dimensional indices,

$$i = \left\lceil \frac{n}{L} \right\rceil, j = n - (i - 1) * L$$

$$\text{for } n = 1, 2, \dots, L, L + 1, \dots, 2L, \dots, (K - 1)L + \dots + KL \quad (2.7)$$

We rewrite $I(X, Y)$ as

$$\begin{aligned} I(X, Y) &= \sum_{i=1}^K \sum_{j=1}^L P_{XY}(x_i, y_j) \log \frac{P_{XY}(x_i, y_j)}{P_1(X=x_i)/P_2(Y=y_j)} \\ &= \sum_{n=1}^{KL} P_{XY}(x_{\lceil n/L \rceil}, y_{n - (\lceil n/L \rceil - 1) * L}) \log \frac{P_{XY}(x_{\lceil n/L \rceil}, y_{n - (\lceil n/L \rceil - 1) * L})}{P_1(X=x_{\lceil n/L \rceil})/P_2(Y=y_{n - (\lceil n/L \rceil - 1) * L})} \end{aligned} \quad (2.8)$$

Since

$$\sum_{n=1}^{KL} P_{XY}(x_{\lceil n/L \rceil}, y_{n - (\lceil n/L \rceil - 1) * L}) = \sum_{i=1}^K \sum_{j=1}^L P_{XY}(x_i, y_j) = 1$$

$$\sum_{n=1}^{KL} P_1(X = x_{\lceil n/L \rceil})/P_2(Y = y_{n - (\lceil n/L \rceil - 1) * L}) = \sum_{i=1}^K \sum_{j=1}^L P_1(X = x_i)/P_2(Y = y_j) = 1 \quad (2.9)$$

We obtain

$$I(X, Y) = \sum_{n=1}^{KL} P_{XY}(X_{[n/L], Y_{n-([n/L]-1)*L}}) \log \frac{P_{XY}(X_{[n/L], Y_{n-([n/L]-1)*L}})}{P_1(X=x_{[n/L]})/P_2(Y=y_{n-([n/L]-1)*L})} \quad (2.10)$$

Claim 2.3 Let X and Y be two discrete random variables. The information exchanged between X and Y meets

$$I(X, Y) \geq 0 \quad (2.11)$$

if and only if X and Y are independent, then with equality.

Proof. To demonstrate the claim 2.2, let's utilise the mappings between two-dimensional indices and one-dimensional indices.

$I(X, Y) \geq 0$ with equality if and only if $P_{XY}(X_{[n/L], Y_{n-([n/L]-1)*L}}) = P_1(X = x_{[n/L]})P_2(Y = y_{n-([n/L]-1)*L})$ for $n = 1, 2, \dots, L, L+1, \dots, 2L, \dots, (K-1)L + \dots + K$ that is, $P_{XY}(x_i, y_j) = P_1(X = x_i)P_2(Y = y_j)$ for $i = 1, \dots, K$ and $j = 1, 2, \dots, L$, or X and Y are separate.

$$I(X, Y) = 0 \quad (2.12)$$

Proof. Suppose the range of X is a constant x and the sample space has only one point ω . Then, $P_1(X = x) = P_1(\omega) = 1$.

For any $j = 1, 2, \dots, L$,

$$P_{XY}(x, y_j) = \sum_{i=1}^1 P_{XY}(x, y_j) = P_2(Y = y_j) = P_1(X = x)/P_2(Y = y_j) \quad (2.13)$$

Thus, X and Y are independent.

Claim 2.4 Let X be discrete random variables with K values. Then

$$0 \leq H(X) \leq \log K \quad (2.12)$$

Proof. Consider two discrete random variables, X and Y. Then, the connections between conditional entropy, entropy, and mutual information are as follows:

$$I(X, Y) = H(X) - H(X/Y) = H(Y) - H(Y/X) = I(Y, X) \quad (2.15)$$

Consider

$$\begin{aligned} I(X, Y) &= \sum_{i=1}^K \sum_{j=1}^L P_{XY}(x_i, y_j) \log \frac{P_{XY}(x_i, y_j)}{P_1(X = x_i)/P_2(Y = y_j)} \\ &= \sum_{i=1}^K \sum_{j=1}^L P_{XY}(x_i, y_j) \log \frac{P_{Y/X}(x_i, y_j)}{P_2(Y = y_j)} = - \sum_{i=1}^K \sum_{j=1}^L P_{XY}(x_i, y_j) \log P_2(Y = y_j) \\ &\quad + \sum_{i=1}^K \sum_{j=1}^L P_{XY}(x_i, y_j) \log P_{Y/X}(x_i, y_j) \\ &= - \sum_{i=1}^K \sum_{j=1}^L P_{XY}(x_i, y_j) \log P_2(Y = y_j) - (- \sum_{i=1}^K \sum_{j=1}^L P_{XY}(x_i, y_j) \log P_{Y/X}(x_i, y_j)) \\ &= - \sum_{j=1}^L \sum_{i=1}^K P_{XY}(x_i, y_j) \log P_2(Y = y_j) - (- \sum_{i=1}^K \sum_{j=1}^L P_{XY}(x_i, y_j) \log P_{Y/X}(x_i, y_j)) \\ &= - \sum_{j=1}^L P_2(Y = y_j) \log P_2(Y = y_j) - (- \sum_{i=1}^K \sum_{j=1}^L P_{XY}(x_i, y_j) \log P_{Y/X}(x_i, y_j)) \\ &= H(Y) - H(Y/X) \end{aligned} \quad (2.16)$$

Claim 2.5 If all the values of X are distinct, then $I(X, Y) = H(Y)$.

Proof. If every value of X is unique, then $K=N$, or the number of distinct values of X equals the number of observations.

we observe that

(1) For all $i = 1, 2, \dots, K$ and $j = 1, 2, \dots, L$, $O_{ij} = 0$ or 1.

(2) For all $i = 1, 2, \dots, K$ and $j = 1, 2, \dots, L$, $p_{ij} = O_{ij}/N = 0$ or $1/N$

(3) for each $j = 1, 2, \dots, L$ since $O_{1j} + O_{2j} + \dots + O_{Kj} = n_j$

Nonzero's O_{ij} or similar n_j nonzero p_{ij} 's, are present.

$$(4) p_{i\cdot} = 1/N, i = 1, 2, \dots, K.$$

With the aforementioned findings and the knowledge that $0 \log 0 = 0$, we have

$$\begin{aligned} I(X, Y) &= \sum_{i=1}^K \sum_{j=1}^L p_{ij} \log \frac{p_{ij}}{p_{i\cdot} p_{\cdot j}} \\ &= \sum_{i=1}^K p_{i1} \log \frac{p_{i1}}{p_{i\cdot} p_{\cdot 1}} + \sum_{i=1}^K p_{i2} \log \frac{p_{i2}}{p_{i\cdot} p_{\cdot 2}} + \dots + \sum_{i=1}^K p_{iL} \log \frac{p_{iL}}{p_{i\cdot} p_{\cdot L}} \\ &= \sum_{p_{i1} \neq 0} \frac{1}{N} \log \frac{1/N}{p_{i1}/N} + \sum_{p_{i2} \neq 0} \frac{1}{N} \log \frac{1/N}{p_{i2}/N} + \dots + \sum_{p_{iL} \neq 0} \frac{1}{N} \log \frac{1/N}{p_{iL}/N} \\ &= \sum_{p_{i1} \neq 0} \frac{1}{N} \log \frac{1}{p_{i1}} + \sum_{p_{i2} \neq 0} \frac{1}{N} \log \frac{1}{p_{i2}} + \dots + \sum_{p_{iL} \neq 0} \frac{1}{N} \log \frac{1}{p_{iL}} \\ &= \frac{n_1}{N} \log \frac{1}{p_1} + \frac{n_2}{N} \log \frac{1}{p_2} + \dots + \frac{n_L}{N} \log \frac{1}{p_L} \\ &= p_1 \log \frac{1}{p_1} + p_2 \log \frac{1}{p_2} + \dots + p_L \log \frac{1}{p_L} \\ &= H(Y) \end{aligned} \quad (2.17)$$

Claim 2.6 Let p and q be two positive numbers less than or equal to 1. Then

$$p \ln \frac{p}{q} \approx (p - q) + \frac{(p-q)^2}{2q} \quad (2.18)$$

The equality holds if and only if $p=q$. Additionally, p is near to q .

Proof. Let's now use $\hat{I}(X, Y)$ to represent $N \times I(X, Y)$. Then, applying claim 2.5, we obtain $2\hat{I}(X, Y) = 2N \sum_{i=1}^K \sum_{j=1}^L p_{ij} \log \frac{p_{ij}}{p_{i\cdot} p_{\cdot j}}$

$$\begin{aligned} &= 2 \sum_{i=1}^K \sum_{j=1}^L O_{ij} \log \frac{O_{ij}/N}{(n_i/N)(n_j/N)} \\ &= 2 \sum_{i=1}^K \sum_{j=1}^L O_{ij} \log \frac{O_{ij}}{n_i n_j / N} \\ &\approx 2 \sum_{i=1}^K \sum_{j=1}^L (O_{ij} - \frac{n_i n_j}{N}) + \sum_{i=1}^K \sum_{j=1}^L \frac{(O_{ij} - \frac{n_i n_j}{N})^2}{n_i n_j / N} \\ &= 2 \sum_{i=1}^K \sum_{j=1}^L O_{ij} - 2 \frac{\sum_i n_i \sum_j n_j}{N} + \sum_{i=1}^K \sum_{j=1}^L \frac{(O_{ij} - \frac{n_i n_j}{N})^2}{n_i n_j / N} \\ &= 2N - 2 \frac{N}{N} + \sum_{i=1}^K \sum_{j=1}^L \frac{(O_{ij} - \frac{n_i n_j}{N})^2}{n_i n_j / N} \\ &= \sum_{i=1}^K \sum_{j=1}^L \frac{(O_{ij} - \frac{n_i n_j}{N})^2}{n_i n_j / N} = \chi^2 \end{aligned} \quad (2.19)$$

[5] states that it has a degree of freedom of $(K-1)(L-1)$ and follows a distribution. Therefore, $2N \times I(X, Y)$ roughly follows a χ^2 distribution with $(K-1)(L-1)$ as its degree of freedom. The well-known Chi-square test for determining whether two random variables are independent is this one. This enables the assignment of a significant level that corresponds to the values of mutual information and $(K-1)(L-1)$ using the Chi-square distribution.

These are the null and alternative hypotheses

H_0 : There is no tie between X and Y ; they are independent.

H_1 : There is a relationship between X and Y , meaning that they are dependent on one other.

The null hypothesis is rejected at the α significance level according to the decision rule if the χ^2 statistic

$$\sum_{i=1}^K \sum_{j=1}^L \frac{(O_{ij} - \frac{n_i n_j}{N})^2}{n_i n_j / N} \approx 2N/I(X, Y) \quad (2.20)$$

is greater than χ^2_{α} , the upper-tail critical value from a Chi-square distribution with $(K-1)(L-1)$ degrees of freedom. That is,

$$\text{Select feature } X \text{ if } I(X, Y) > \frac{\kappa_U^2}{2N} \quad (2.21)$$

Conclusions

Using random variables with distinct probability spaces, we have provided a unified definition of mutual information in this study. Our proposal is to use the marginal probabilities to define the joint distribution of two random variables. Different joint distributions will yield different mutual information values according to our revised definition of mutual information. We first established a few characteristics of the newly defined mutual information, and then we suggested a machine learning calculation approach. Lastly, we used credit rating with our newly specified mutual information.

References

1. Tourassi GD, Frederick ED, Markey MK, Floyd CE Jr. Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Med Phys*. 2001;28(12):2394–2402.
2. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*. 2003;3:1157–1182.
3. Navot A. On the role of feature selection in machine learning. [PhD thesis]. Hebrew University; c2006.
4. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27(3):379–423.
5. Kullback S. *Information Theory and Statistics*. New York, NY: John Wiley & Sons; c1959.
6. Pinsker MS. *Information and Information Stability of Random Variables and Processes*. Academy of Science, USSR; c1960. (English translation by A. Feinstein, Holden-Day, San Francisco, USA, 1964).
7. Ash RB. *Information Theory*. New York, NY: Interscience Publishers; c1965.
8. Cover TM, Thomas JA. *Elements of Information Theory*. 2nd ed. New York, NY: John Wiley & Sons; c2006.
9. Fano RM. *Transmission of Information*. Cambridge, MA: MIT Press; New York, NY: John Wiley & Sons; c1961.
10. Abramson N. *Information Theory and Coding*. New York, NY: McGraw-Hill; c1963.
11. Gallager RG. *Information Theory and Reliable Communication*. New York, NY: John Wiley & Sons; c1968.
12. Ash RB, Doleans-Dade CA. *Probability & Measure Theory*. 2nd ed. San Diego, CA: Academic Press; c2000.
13. Braga I. A constructive density-ratio approach to mutual information estimation: experiments in feature selection. *J Inf Data Manag*. 2014;5(1):134–143.
14. Verma RK. Information radius via Verma information measure in intuitionistic fuzzy environment. *Int J Math Res*. 2023;15(1):1-8.
15. Verma RK. Modified version of Verma measures of information and their kinship with past information measures. *Int J Pure Appl Math Sci*. 2023;16(1):17–24.
16. Verma RK. Obvious disparities on optimal guesswork wiretapper moments under mismatch related to non-Shannon cypher system. *Asian J Prob Stat*. 2023;24(4):17–22. DOI:10.9734/AJPAS/2023/v24i4531.
17. Verma RK. On optimal channel capacity theorems via Verma information measure with two-sided input in noisy state. *Asian J Prob Stat*. 2023;22(2):1–7. DOI:10.9734/AJPAS/2023/v22i2478.